



2008 Report to Congress

Data Mining: Technology and Policy
The DHS Privacy Office

December 2008



Homeland
Security

Report Documentation Page			Form Approved OMB No. 0704-0188		
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE DEC 2008		2. REPORT TYPE		3. DATES COVERED 00-00-2008 to 00-00-2008	
4. TITLE AND SUBTITLE 2008 Report to Congress, Data Mining: Technology and Policy The DHS Privacy Office				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Department of Homeland Security, Washington, DC				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 47	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			



2008 Report to Congress
Data Mining: Technology and Policy

Respectfully submitted,

Hugo Teufel III
Chief Privacy Officer
U.S. Department of Homeland Security
Washington, DC

December 2008

TABLE OF CONTENTS

1. Executive Summary	1
2. Introduction	3
3. Background	5
3.1. Data Mining and the Privacy Office Compliance Process	5
3.2. Data Mining Reporting Act Requirements	8
4. Reporting	10
4.1. Automated Targeting System (ATS)	10
4.1.1. General Program Description	10
4.1.2. ATS-Inbound and ATS-Outbound Modules (Cargo Analytics)	12
4.1.3. ATS-Passenger Module	15
4.1.4. ATS Privacy Impact and Privacy Protections	18
4.2. Data Analysis and Research for Trade Transparency System (DARTTS)	20
4.2.1. Program Description	20
4.2.2. Technology and Methodology	21
4.2.3. Data Sources	23
4.2.4. Efficacy	24
4.2.5. Laws and Regulations	24
4.2.6. Privacy Impact and Privacy Protections	25
4.3. Freight Assessment System (FAS)	27
4.3.1. Program Description	27
4.3.2. Technology and Methodology	28
4.3.3. Data Sources	28
4.3.4. Efficacy	29
4.3.5. Laws and Regulations	29
4.3.6. Privacy Impact and Privacy Protections	29
5. Public Workshop on Privacy Protections for Government Data Mining	30
5.1. Defining “Data Mining”	31
5.2. Panel on How Government Data Mining Impacts Privacy	32
5.3. Panel on Validating Data Mining Models and Results	33
5.4. Panel on Technologies for Privacy-Protective Data Mining	34
5.5. Panel on Auditing and Enforcing Privacy Controls for Data Mining	35
5.6. Panel on Privacy Policies and Best Practices for Government Data Mining	35
6. Privacy Principles for DHS S&T Research Projects	37
6.1. Background	37
6.2. Privacy Principles for DHS S&T Research Projects	37
7. Conclusion	39
8. Appendix	41

1. Executive Summary

The Department of Homeland Security Privacy Office is pleased to provide to the Congress its 2008 report *Data Mining: Technology and Policy*. The Privacy Office has prepared this report to the Congress pursuant to the Department's obligations under Section 804 of the *Implementing the Recommendations of the 9/11 Commission Act of 2007*, entitled the *Federal Agency Data Mining Reporting Act of 2007*.¹ This report discusses activities currently deployed or under development in the Department that meet the Act's definition of data mining, and provides the information set out in the Act's reporting requirements for data mining programs. It also provides a summary of the Privacy Office's public workshop, "*Implementing Privacy Protections in Government Data Mining*," which was held on July 24-25, 2008. Finally, this report presents new privacy principles for research projects conducted by the DHS Science and Technology Directorate (S&T), the Department's primary research and development arm. The *Principles for Implementing Privacy Protections in S&T Research Projects* were developed jointly by the Privacy Office and S&T and will be implemented in all new S&T research projects, including those that involve data mining.

The Privacy Office has identified three Department programs that currently engage in data mining, as defined by the *Federal Agency Data Mining Reporting Act*: the Automated Targeting System (ATS) Inbound, Outbound, and Passenger modules, which are administered by U.S. Customs and Border Protection (CBP); the Data Analysis and Research for Trade Transparency System (DARTTS), which is administered by Immigration and Customs Enforcement (ICE); and the Freight Assessment System (FAS), which is administered by the Transportation Security Administration (TSA). While each of these programs engages to some extent in data mining, none uses data mining to make unevaluated automated decisions about individuals, *i.e.*, none of these programs makes decisions about individuals solely on the basis of data mining results. In all cases, DHS employees conduct investigations to verify (or disprove) the results of data mining, and then bring their own judgment and experience to bear in making determinations about individuals initially identified through data mining activities.

For each program enumerated below, this report describes the program's purpose and methodology, the technology employed, the legal authority for the program, and the sources of the data the program uses. Each program description also includes an analysis of the program's efficacy and concludes with a discussion of the program's impacts on individual privacy and the protections in place to address those impacts.

¹ Pub. L. 110-53, 121 Stat. 266, Section 804.

As described more fully below, the Privacy Office's compliance process requires programs using personally identifiable information (PII) to have completed federally-mandated privacy documentation, consisting of a Privacy Impact Assessment (PIA)² and a System of Records Notice (SORN).³ The Privacy Office has worked closely with the programs discussed in this report to complete the required privacy compliance documentation. Programs that use PII have issued both PIAs and SORNs.

The report section on DHS data mining activities is followed by a summary of the Privacy Office's public workshop on *Implementing Privacy Protections in Government Data Mining*, which brought academics, government researchers, policy and technology experts, and privacy advocates together to discuss the privacy issues associated with government activities that involve data mining and to identify ways to address those issues. Participants described the actual and potential impacts of government data mining on individuals and on society. They explored methods of validating the accuracy and effectiveness of data mining models and results, and they discussed anonymization tools and audit controls that could be implemented to enhance privacy protection in data mining projects. Building upon the current consensus among many government and private-sector groups, academics, privacy advocates, security experts, and others who have studied the privacy issues associated with data mining, participants identified key elements of best practices for engaging in data mining activities in a manner that both protects privacy and furthers the Department's mission.

As an outgrowth of the public workshop, the Privacy Office and S&T began working to identify privacy principles that could be embedded in S&T's research and development projects involving data mining. As this effort progressed, S&T and the Privacy Office determined to broaden the analysis to include all privacy-sensitive S&T research. This collaboration has led to a set of *Principles for Implementing Privacy Protections in S&T Research* (Principles), which S&T has agreed will govern new research performed at S&T laboratories, S&T-sponsored research conducted in cooperation with other Federal government entities, and research conducted by external performers under a contract with S&T.

² Section 208 of the E-Government Act of 2002, Pub. L. 107-347 (December 17, 2002), specifically requires that when a Federal agency develops a new information technology the agency must conduct an analysis of that technology's impact on privacy. In addition, Section 222 of the Homeland Security Act enhances the Privacy Office's ability to conduct PIAs on the use of technology at the Department, by requiring the Chief Privacy Officer to ensure that "the use of technologies sustains, and does not erode, privacy protections relating to the use, collection, and disclosure of personal information." 6 U.S.C. § 142(1).

³ The Privacy Act of 1974 requires Federal agencies to implement fair information principles when handling collections of personal information. 5 U.S.C. §552a, as amended. If an agency maintains a System of Records, as defined in 5 U.S.C. §552a(a)(5), then that agency must provide notice regarding that collection, informing the public about, *inter alia*, the collection's purpose, who is affected by the collection, what types of information is contained in the collection, and how the information will be shared outside the agency. 5 U.S.C. § 552a(e)(4).

Consistent with privacy implementation generally at the Department, and with the best practices identified by workshop participants, the Principles reflect the Fair Information Practice Principles of Transparency, Individual Participation, Purpose Specification, Data Minimization, Use Limitation, Data Quality and Integrity, Security, and Accountability and Auditing. The Principles also include a *Privacy Assessment Principle*, which reinforces that an assessment of privacy impacts, conducted jointly by S&T and the Privacy Office, will be an integral part of the development and implementation of any S&T research project that is privacy-sensitive and/or involves or impacts PII. With development of the Principles completed, the Privacy Office and S&T will now turn to creating an implementation plan for applying the Principles to new S&T research projects.

2. Introduction

The Privacy Office operates under the direction of the Chief Privacy Officer, who is appointed by and reports directly to the Secretary of the Department of Homeland Security (“Department” or DHS). The Privacy Office serves to implement Section 222 of the Homeland Security Act of 2002, as amended,⁴ and has programmatic responsibilities involving the Privacy Act of 1974,⁵ the Freedom of Information Act (FOIA),⁶ the privacy provisions of the E-Government Act of 2002,⁷ and DHS policies that protect individual privacy associated with the collection, use, and disclosure of personally identifiable information (PII). Section 222 of the Homeland Security Act of 2002 requires the Chief Privacy Officer to assume primary responsibility for privacy policy within the Department.⁸ In addition, Section 802 of the *Implementing the Recommendations of the 9/11 Commission Act of 2007* (“9/11 Commission Act”) codifies the Chief Privacy Officer’s authority to investigate and/or report on DHS programs and operations with respect to privacy.⁹

The Privacy Office has published two previous reports on DHS data mining activities, using varying definitions of data mining. The Privacy Office published its first such report,

⁴ Section 222 of the Homeland Security Act of 2002, as amended by Section 8305 of the Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. 108-458 (December 17, 2004), 6 U.S.C. § 142.

⁵ 5 U.S.C. §552a.

⁶ 5 U.S.C. § 552.

⁷ Pub. L. 107-347, 116 Stat. 2899, § 208; 44 U.S.C. § 101 et seq.

⁸ 6 U.S.C. §142(a).

⁹ Pub. L. 110-53, 121 Stat. 266. For a complete description of the Chief Privacy Officer’s statutory duties and responsibilities, see Dept. of Homeland Security, *Privacy Office Annual Report to Congress July 2007-July 2008*, at 1-2(available on the Privacy Office website at http://www.dhs.gov/xinfoshare/publications/editorial_0514.shtm).

entitled *Data Mining Report: DHS Privacy Office Response to House Report 108-774*, on July 6, 2006 (“2006 Report”). The 2006 Report was prepared pursuant to the requirements of House Report 108-774 – *Making Appropriations for the Department of Homeland Security for the Fiscal Year Ending September 30, 2005, and for Other Purposes*,¹⁰ The 2006 Report used a definition of data mining derived from definitions used by the Congressional Research Service and the Government Accountability Office.¹¹ It provided a description of the process of data mining, and set out the privacy concerns raised by the use of data mining technologies for homeland security. It identified specific data mining activities and programs at DHS and provided information related to their purpose, data sources, and deployment dates. The 2006 Report also described the policies, procedures, and guidance that applied to each data mining activity identified. Looking forward, the 2006 Report made a number of recommendations aimed specifically at addressing privacy concerns DHS data mining activities might raise.¹²

On July 6, 2007, the Privacy Office published its second report on Department data mining activities, *2007 Data Mining Report: DHS Privacy Office Response to House Report 109-699* (“2007 Report”).¹³ The 2007 Report described DHS activities that met the definition of data mining required by H.R. 109-699.¹⁴ It discussed the Privacy Office’s initial efforts to promote implementation of the 2006 Report’s recommendations in DHS data mining programs. The Privacy Office also stated its interest in holding a public workshop to explore appropriate privacy protections for DHS data mining activities, including the use of anonymization tools.¹⁵ On February 11, 2008, the Privacy Office issued a *Letter Report Pursuant to Section 804 of the Implementing Recommendations of the 9/11 Commission Act of 2007* (“Letter Report”), which included a brief discussion of relevant DHS data mining activities, with the understanding that a comprehensive report would follow. The Letter Report reiterated the Office’s interest in holding a public workshop to discuss the privacy impacts of government data mining, to identify ways of

¹⁰ Conference Report on H.R. 4567 (enacted as Pub. L. 108-334, 118 Stat.1298, Oct. 18, 2004).

¹¹ 2006 Report at 6-8.

¹² The 2006 Report is available on the Privacy Office web site at http://www.dhs.gov/xinfoshare/publications/editorial_0514.shtm.

¹³ The Privacy Office prepared its second report pursuant to the requirements of House Report No. 109-699 – *Making Appropriations for the Department of Homeland Security for the Fiscal Year Ending September 30, 2007, and for Other Purposes*. H. Conf. Rept. 109-699, Making Appropriations for the Department of Homeland Security for the Fiscal Year Ending September 30, 2007, and for Other Purposes, Sept. 28, 2006, H7784, at H7815 (Conference Report on H.R. 5441).

¹⁴ 2007 Report at 6-8.

¹⁵ 2007 Report at 34. The 2007 Report is available on the Privacy Office web site at http://www.dhs.gov/xinfoshare/publications/editorial_0514.shtm,

validating data mining models, and to showcase technology tools that could both enhance privacy and support data mining research.¹⁶

The Privacy Office is providing this 2008 report to the Congress pursuant to Section 804 of the 9/11 Commission Act,¹⁷ entitled *The Federal Agency Data Mining Reporting Act of 2007* (*Data Mining Reporting Act* or Act). The Act provides a new definition of “data mining” and requires the Department to report annually to the Congress. This report is the Privacy Office’s first comprehensive submission to the Congress under the Act. It discusses DHS programs that satisfy the Act’s definition of “data mining” in light of the Act’s content-reporting requirements. It summarizes the Privacy Office’s public workshop, *Implementing Privacy Protections in Government Data Mining*, which was held on July 24-25, 2008. Finally, this report presents newly-announced privacy principles, developed jointly by the Privacy Office and the DHS Science and Technology Directorate (S&T), for research activities conducted by S&T, including those that involve data mining.

3. Background

3.1. Data Mining and the Privacy Office Compliance Process

The *Homeland Security Act of 2002* expressly authorizes the Department to use data mining, among other analytical tools, in furtherance of its mission.¹⁸ DHS exercises this authority to engage in data mining in the programs discussed in this report. Three Federal laws form the foundation for privacy protections for data mining activities at the Department: Section 222 of the *Homeland Security Act of 2002*,¹⁹ the *Privacy Act of 1974*,²⁰ and the *E-Government Act of 2002*.²¹ Section 222 of the *Homeland Security Act of 2002* states that the DHS Chief Privacy Officer is responsible for “assuring that the use of technologies sustains, and does not erode, privacy protections relating to the use, collection, and disclosure of personal

¹⁶ The Letter Report is available on the Privacy Office’s website at http://www.dhs.gov/xinfoshare/publications/editorial_0514.shtm.

¹⁷ Pub. L. No. 110-53, 121 Stat. 266.

¹⁸ The Act states that, “[s]ubject to the direction and control of the Secretary, the responsibilities of the Under Secretary for Information Analysis and Infrastructure Protection, shall be as follows . . . To establish and utilize, in conjunction with the chief information officer of the Department, a secure communications and information technology infrastructure, including data mining and other advanced analytical tools, in order to access, receive, and analyze data and information in furtherance of the responsibilities under this section, and to disseminate information acquired and analyzed by the Department, as appropriate.” 6 U.S.C §121(d)(14).

¹⁹ 6 U.S.C. § 142.

²⁰ 5 U.S.C. § 552a.

²¹ Pub. L. 107-347, 116 Stat. 2899, § 208; 44 U.S.C. § 101 *et seq.* .

information.”²² The Chief Privacy Officer is also responsible for “assuring that personal information contained in Privacy Act systems of records is handled in full compliance with fair information practices as set out in the *Privacy Act of 1974*.”²³

The Privacy Office uses a number of tools to ensure that all departmental activities, including those that involve data mining, meet these requirements. First, the Privacy Act, among other requirements, mandates that agencies publish a notice when PII is maintained in a system of records.²⁴ The System of Records Notice (SORN) is published in the Federal Register and identifies the purpose for the system of records, the categories of individuals in the system, what categories of information are maintained about the individuals, and how the agency discloses the information to other agencies (routine uses).²⁵ The SORN also provides public notice regarding the available mechanisms to exercise the rights granted through the Privacy Act to access and correct the PII an agency maintains.²⁶

Second, Section 208 of the *E-Government Act of 2002* requires all Federal agencies to conduct Privacy Impact Assessments (“PIA”) for all new technology that collects, maintains, or disseminates PII.²⁷ Office of Management and Budget (OMB) guidance for implementing the privacy provisions of the E-Government Act extends this requirement to technology that predates the E-Government Act, if that technology has since undergone a significant change.²⁸ The E-Government Act does not require PIAs on national security systems or systems containing information about Federal employees and contractors;²⁹ as a policy matter, however, the Privacy Office requires all information technology systems, including classified systems, to conduct PIAs.³⁰ Nonetheless, classified systems may be exempted from the requirement to publish a

²² 6 U.S.C. § 142(1).

²³ 6 U.S.C. § 142(2).

²⁴ 5 U.S.C. § 552a(e)(4) (“[S]ubject to the provisions of paragraph (11) of this subsection, [each agency that maintains a system of records shall] publish in the Federal Register upon establishment or revision a notice of the existence and character of the system of records....”). The term “system of records” means “a group of any records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual” 5 U.S.C. § 552a(a)(5).

²⁵ 5 U.S.C. § 552a(e)(4)

²⁶ *Id.*

²⁷ Pub. L. 107-347, 116 Stat. 2899, § 208(b), 44 U.S.C. § 3501 note.

²⁸ OMB *Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002*, M-03-22, Sept. 26, 2003.

²⁹ The Privacy Office requires a PIA for Federal employee and contractor systems that are deployed throughout DHS.

³⁰ *See* 6 U.S.C. § 142(1).

PIA. As part of its overall compliance program, the Privacy Office identifies programs that engage in data mining through several different processes. First, the Privacy Office reviews all OMB-300 budget submissions, to learn of programs or systems that use PII and to determine whether they address privacy appropriately. Second, the Privacy Office created the Privacy Threshold Analysis (PTA) document, a form used for all information technology systems that are going through the certification and accreditation (C&A) process required under the Federal Information Security Management Act (FISMA),³¹ to determine whether they maintain PII. The PTA assists the Privacy Office in identifying those IT systems that use PII and, of those, which must conduct PIAs and whether a SORN covers the system. During the PIA process, the Privacy Office identifies those systems that need a new or updated SORN as required by the Privacy Act. Through the PIA process, specific questions related to analytical use of data are asked, to identify and mitigate any privacy risks from the use of such technology.

In addition, the Privacy Office reviews technology investment proposals that the DHS Enterprise Architecture Center of Excellence (EACOE) and Integrated Project Team (IPT) process, to ensure that DHS investments in technology include a specific review for compliance with privacy protection requirements. Through these activities, the Privacy Office compliance process provides a number of opportunities to learn about proposed data mining activities and to engage program managers in discussions about potential privacy issues.

All of the Privacy Office's compliance tools apply the Fair Information Practice Principles (FIPPs), the fundamental framework for privacy implementation at the Department: Transparency, Individual Participation, Purpose Specification, Data Minimization, Use Limitation, Data Quality and Integrity, Security, and Accountability and Auditing. These principles are also reflected in agency SORNs, which are issued pursuant to the *Privacy Act of 1974* and published in the Federal Register.

The Privacy Office has worked closely with the relevant DHS components to complete the privacy compliance documentation required for each of the programs described in this Report. As discussed more fully below, a PTA has been completed for all of these programs; programs that use PII have issued both PIAs and SORNs.

³¹ The Federal Information Security Management Act is included under Title III of the *E-Government Act* (Pub. L. 107-347) and is codified at 44 U.S.C. § 3541 *et seq.*

3.2. Data Mining Reporting Act Requirements

As noted above, this report is provided to the Congress pursuant to the *Data Mining Reporting Act*, which defines “data mining” as:

a program involving pattern-based queries, searches, or other analyses of 1 or more electronic databases, where—

(A) a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals;

(B) the queries, searches, or other analyses are not subject-based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases; and

(C) the purpose of the queries, searches, or other analyses is not solely—

(i) the detection of fraud, waste, or abuse in a Government agency or program; or

(ii) the security of a Government computer system.³²

The Act excludes queries, searches or analyses that are conducted solely in electronic databases of publicly-available information: telephone directories, news reporting services, databases of legal and administrative rulings, and other databases and services providing public information without a fee.³³

It is important to note two key aspects of the Act’s definition of “data mining.” First, the definition is limited to *pattern-based* electronic searches, queries or analyses; activities that use only personally identifying information, or other terms specific to individuals (*e.g.*, a license plate number or vessel registration number), as search terms are excluded from the definition. Second, the definition is limited to searches, queries or analyses that are conducted for the purpose of identifying predictive patterns or anomalies that are indicative of terrorist or criminal

³² Pub. L. 110-53, 121 Stat. 266, Section 804(b)(1).

³³ *Id.* at Section 804(b)(2).

activity by an individual or individuals. Research in electronic databases that produces only a summary of historical trends is not “data mining” under the Act.

The Act requires the Department to provide the Congress the following information about each program or activity that meets the Act’s definition of “data mining:”

A thorough description of the data mining activity, its goals, and, where appropriate, the target dates for the deployment of the data mining activity.

A thorough description of the data mining technology that is being used or will be used, including the basis for determining whether a particular pattern or anomaly is indicative of terrorist or criminal activity.

A thorough description of the data sources that are being or will be used.

An assessment of the efficacy or likely efficacy of the data mining activity in providing accurate information consistent with and valuable to the stated goals and plans for the use or development of the data mining activity.

An assessment of the impact or likely impact of the implementation of the data mining activity on the privacy and civil liberties of individuals, including a thorough description of the actions that are being taken or will be taken with regard to the property, privacy, or other rights or privileges of any individual or individuals as a result of the implementation of the data mining activity.

A list and analysis of the laws and regulations that govern the information being or to be collected, reviewed, gathered, analyzed, or used in conjunction with the data mining activity, to the extent applicable in the context of the data mining activity.

A thorough discussion of the policies, procedures, and guidelines that are in place or that are to be developed and applied in the use of such data mining activity in order to protect the privacy and due process rights of individuals, such as redress procedures, and ensure that only accurate and complete information is collected, reviewed, gathered, analyzed, or used, and guard against any harmful consequences of potential inaccuracies.³⁴

The Privacy Office addresses these reporting requirements for each of the DHS programs included in this report.

³⁴ *Id.* at Section 804(c)(2).

4. Reporting

The Privacy Office has identified three DHS programs that currently engage in data mining, as defined by the Data Mining Reporting Act:

- The Automated Targeting System (ATS) Inbound, Outbound, and Passenger modules(administered by U.S. Customs and Border Protection (CBP));
- The Data Analysis and Research for Trade Transparency System (DARTTS) (administered by Immigration and Customs Enforcement (ICE)); and
- The Freight Assessment System (FAS) (administered by the Transportation Security Administration (TSA)).³⁵

The descriptions that follow address the Data Mining Reporting Act's reporting requirements for each of these programs. It is important to note at the outset that, while each of the following programs engages to some extent in data mining, no program makes decisions about individuals solely on the basis of data mining results. In all cases, DHS employees conduct investigations to verify (or disprove) the results of data mining, and then bring their own judgment and experience to bear in making determinations about individuals initially identified through data mining activities.

4.1. Automated Targeting System (ATS)

4.1.1. General Program Description

CBP has developed the Automated Targeting System (ATS), an intranet-based enforcement and decision support tool that is the cornerstone for all CBP targeting efforts. ATS compares traveler, cargo, and conveyance information against intelligence and other enforcement data by incorporating risk-based targeting scenarios and assessments. CBP uses ATS to improve the collection, use, analysis, and dissemination of information that is gathered for the primary purpose of targeting, identifying, and preventing potential terrorists and terrorist weapons from entering the United States. CBP also uses ATS to identify other violations of U.S. laws that are enforced by CBP. In this way, ATS allows CBP officers charged with enforcing U.S. law and preventing terrorism and other crime to focus their efforts on travelers, conveyances, and cargo shipments that most warrant greater scrutiny. ATS standardizes names, addresses, conveyance

³⁵ ICE's Law Enforcement Intelligence Fusion System ("IFS") (formerly known as the ICE Network Law Enforcement Analysis Data System (NETLEADS)), which was described in the 2007 Report, no longer engages in data mining, as defined in the Data Mining Reporting Act. All electronic searches, queries and analyses currently conducted by IFS are subject-based and use personal identifiers or other inputs associated with specific individual(s) as search terms.

names, and similar data so these data elements can be more easily associated with other business data and personal information to form a more complete picture of a traveler, import, or export in context with previous behavior of the parties involved. Traveler, conveyance, and shipment data are processed through ATS and are subject to a real-time, rules-based evaluation.

ATS consists of six modules that focus on exports, imports, passengers and crew (airline passengers and crew on international flights, and passengers and crew on sea carriers), private vehicles crossing at land borders, and import trends over time. This report discusses three of these modules: ATS-Inbound, ATS-Outbound, and ATS-Passenger (ATS-P). The remaining modules do not involve data mining as defined by the *Data Mining Reporting Act*.³⁶

As a legacy organization of CBP, the U.S. Customs Service traditionally employed computerized screening tools to target potentially high-risk cargo entering, exiting, and transiting the United States. ATS was originally designed as a rules-based program to identify such cargo; it did not apply to travelers. ATS-Inbound and ATS-Outbound became operational in 1997. ATS-P became operational in 1999 and is critically important to CBP's mission. ATS-P allows CBP officers to determine whether a variety of potential risk indicators exist for travelers and/or their itineraries that may warrant additional scrutiny. ATS-P maintains Passenger Name Record (PNR) data, which is data provided to airlines and travel agents by or on behalf of air passengers seeking to book travel. CBP began receiving PNR data voluntarily from certain air carriers in 1997. Currently, CBP collects this information to the extent collected by carriers in connection with a flight into or out of the United States, as part of its border enforcement mission and pursuant to the *Aviation and Transportation Security Act of 2001* (ATSA).³⁷

ATS receives various data in real time from the following CBP mainframe systems: the Automated Commercial System (ACS), the Automated Manifest System (AMS), the DHS Advance Passenger Information System (APIS), the Automated Export System (AES), the Automated Commercial Environment (ACE), and the Treasury Enforcement Communications System (TECS). TECS includes information from the Federal Bureau of Investigation Terrorist Screening Center's³⁸ Terrorist Screening Database (TSDB) and other government databases

³⁶ These modules are: ATS-Land (which provides targeting capability for private vehicles arriving by land); ATS-Trend Analysis and Analytical Selectivity Program (ATS-TAP) (which provides trend analysis of historical international trade statistics to identify anomalous activity in aggregate); and ATS-International, which is being developed to support collaborative efforts with foreign customs administrations.

³⁷ Pub. L. 107-71 (2001), as codified at 49 U.S.C. § 44909 (implementing regulations at 19 CFR § 122.49d).

³⁸ The TSC is an entity established by the Attorney General in coordination with the Secretary of State, the Secretary of Homeland Security, the Director of the Central Intelligence Agency, the Secretary of the Treasury, and the Secretary of Defense. The Attorney General, acting through the Director of the FBI, established the TSC in support of Homeland Security Presidential Directive 6 (HSPD-6), dated September 16, 2003, which required the Attorney General to establish an organization to consolidate the Federal Government's approach to terrorism

regarding individuals with outstanding wants and warrants and other high-risk individuals and entities. ATS collects certain data directly from air carriers in the form of PNR. ATS also collects data from foreign governments and certain express consignment services in conjunction with specific cooperative programs. ATS accesses data from these sources, which collectively include: electronically filed bills of lading, entries, and entry summaries for cargo imports; shippers' export declarations and transportation bookings and bills for cargo exports; manifests for arriving and departing passengers; land-border crossing and referral records for vehicles crossing the border; airline reservation data; nonimmigrant entry records; and records from secondary referrals, incident logs, suspect and violator indices, seizures, and information from the TSDB and other government databases regarding individuals with outstanding wants and warrants and other high-risk entities. Finally, ATS uses data from Dun & Bradstreet, a commercially available data source, to assist with company identification through name and address matching.

In addition to providing a risk-based assessment system, ATS provides a graphical user interface (GUI) for many of the underlying legacy systems from which ATS pulls information. This interface improves the user experience by providing the same functionality in a more rigidly controlled access environment than the underlying system. Access to this functionality of ATS uses existing technical security and privacy safeguards associated with the underlying systems.

A large number of rules are included in the ATS modules, which encapsulate sophisticated concepts of business activity that help identify suspicious or unusual behavior. The ATS rules are constantly evolving to both meet new threats and refine existing rules. ATS applies the same methodology to all individuals to preclude any possibility of disparate treatment of individuals or groups. ATS is consistent in its evaluation of risk associated with individuals and is used to support the overall CBP law enforcement mission.

4.1.2. ATS-Inbound and ATS-Outbound Modules (Cargo Analytics)

4.1.2.1. Program Description

ATS-Inbound is available to CBP officers at all major ports (air/land/sea/rail) throughout the United States, and also assists CBP personnel in the Container Security Initiative (CSI) decision-making process. ATS-Inbound assists CBP officers in identifying inbound cargo shipments that pose a high risk of containing weapons of mass effect, illegal narcotics, or other contraband, and in selecting that cargo for intensive examination.

screening and provide for the appropriate and lawful use of terrorist information in screening processes. The TSC maintains the Federal Government's consolidated terrorist watch list, known as the TSDB.

ATS-Outbound aids CBP officers in identifying exports that pose a high risk of containing goods requiring specific export licenses, illegal narcotics, smuggled currency, stolen vehicles or other contraband, or exports that may otherwise be in violation of U.S. law. ATS-Outbound sorts Electronic Export Information (EEI) (formerly referred to as the Shippers' Export Declaration (SED)) data extracted from AES, compares it to a set of rules, and evaluates it in a comprehensive fashion. This information assists CBP officers in targeting and/or identifying exports that pose potential aviation safety and security risks (*e.g.*, hazardous materials) or may be otherwise exported in violation of U.S. law.

ATS-Inbound and ATS-Outbound look at data related to cargo in real time and engage in data mining to provide decision support analysis for targeting of cargo for suspicious activity. The cargo analysis provided by ATS is intended to add automated anomaly detection to CBP's existing targeting capabilities, to enhance screening of cargo prior to its entry into the United States.

4.1.2.2. Technology and Methodology

ATS-Inbound and ATS-Outbound do not collect information directly from individuals. The data used in the development and testing of ATS-Inbound and ATS-Outbound screening technology is taken from bills of lading and shipping manifest data provided by vendors to CBP as part of the existing cargo screening process. The results of queries, searches, and analyses conducted in the ATS-Inbound and ATS-Outbound system are used to identify anomalous business behavior, data inconsistencies, abnormal business patterns, and suspicious business activity generally. No decisions about individuals are made solely on the basis of these results.

The *Security and Accountability for Every Port Act of 2006 (SAFE Port Act)* requires ATS to use or investigate the use of advanced algorithms in support of its mission.³⁹ To that end, ATS has established an Advanced Targeting Initiative, which includes plans for development of data mining, machine learning,⁴⁰ and other analytic techniques during the period from FY09 to FY12, for use in ATS-Inbound and ATS-Outbound. Development will take place in iterative phases; the various iterations will be deployed to a select user population, which will test the new functionality. The Advanced Targeting Initiative is being undertaken in tandem with ATS' maintenance and operation of the ATS-Inbound and ATS-Outbound system. The design and tool-selection processes for data mining, pattern recognition, and machine learning

³⁹ Pub. L. 109-347, October 11, 2006.

⁴⁰ Machine learning is concerned with the design and development of algorithms and techniques that allow computers to "learn." The major focus of machine learning research is to extract information from data automatically, using computational and statistical methods. This extracted information may then be generalized into rules and patterns.

techniques in development in the Advanced Targeting Initiative are under consideration and have yet to be finalized.

4.1.2.3. Data Sources

As noted above, ATS-Inbound and ATS-Outbound do not collect information directly from individuals. The information maintained in ATS is either collected from private entities providing data in accordance with U.S. legal requirements (*e.g.*, sea, rail and air manifests) or is created by ATS as part of its risk assessments and associated rules.

ATS-Inbound and ATS-Outbound use the information in ATS source databases to gather information about importers and exporters, cargo, and conveyances used to facilitate the importation of cargo into and the exportation of cargo out of the United States. This information includes PII concerning individuals associated with imported and exported cargo (*e.g.*, brokers, carriers, shippers, buyers, sellers, exporters, freight forwarders and crew). ATS-Inbound receives data pertaining to entries and manifests from ACS and ACE, and processes it against a variety of rules to make a rapid, automated assessment of the risk of each import.⁴¹ ATS-Outbound uses EEI data that exporters file electronically with AES, export manifest data from AES, export airway bills of lading, and census export data from U.S. Department of Commerce, to assist in formulating risk assessments for cargo bound for destinations outside the United States.

CBP uses commercial off-the-shelf (COTS) software tools to graphically present entity-related information that may represent terrorist or criminal activity, to discover non-obvious relationships across cargo data, to retrieve information from ATS source systems to expose unknown or anomalous activity, and to conduct statistical modeling of cargo-related activities as another approach to detecting anomalous behavior. CBP also uses a custom-designed company resolution application to resolve ambiguities in trade entity identification related to inbound and outbound cargo.

⁴¹ ATS-Inbound collects information regarding individuals in connection with the following items including, but not limited to: Sea/Rail Manifests from AMS; Cargo Selectivity Entries and Entry Summaries from the Automated Broker Interface (ABI), a component of ACS; Air Manifests (bills of lading) from AMS; Express Consignment Services (bills of lading); CCRA Manifests (bills of lading from Canada Customs and Revenue (CCRA)); CBP Automated Forms Entry Systems (CAFES) CBP Form 7512; QP Manifest Inbound (bills of lading) from AMS; Truck Manifests from ACE; Inbound Data (bills of lading) from AMS; entries subject to Food and Drug Administration (FDA) Prior Notice (PN) requirements from ACS; and Census Import Data from the U.S. Department of Commerce.

4.1.2.4. Efficacy

Based upon the results of testing and operations in the field, ATS-Inbound and ATS-Outbound have proved to be effective means of identifying suspicious cargo that requires further investigation by CBP officers. The results of ATS-Inbound and ATS-Outbound analyses identifying cargo as suspicious have been regularly corroborated by physical searches of the cargo so identified.

The goal of the Advanced Targeting Initiative is to enhance CBP officers' ability to identify entities such as organizations, cargo, vehicles, and conveyances with a possible association to terrorism. By their very nature, the results produced by technologies used in the Advanced Targeting Initiative may be only speculative or inferential; they may only provide leads for further investigation rather than a definitive statement. The program finds it valuable to be able to very quickly produce useful leads gleaned from masses of information. Leads resulting in a positive, factual determination obtained through further investigation and physical inspections of cargo demonstrate the efficacy of these technologies.

4.1.2.5. Laws and Regulations

There are numerous customs and immigration authorities authorizing the collection of data regarding the import and export of cargo as well as the entry and exit of conveyances.⁴² Additionally, ATS-Outbound and ATS-Inbound support functions mandated by Title VII of Public Law 104-208 (*1996 Omnibus Consolidated Appropriations Act for FY 1997*), which provides funding for counter-terrorism and drug law enforcement. ATS-Outbound also supports functions arising from the *Anti-Terrorism Act of 1987*⁴³ and the *1996 Clinger-Cohen Act*.⁴⁴ The risk assessments for cargo are also mandated under Section 203 of the *SAFE Port Act*.

4.1.3. ATS-Passenger Module

4.1.3.1. Program Description

ATS-Passenger (ATS-P) is a custom-designed system used at U.S. ports of entry, particularly those receiving international flights and voyages, to evaluate passengers and crewmembers prior to arrival or departure. ATS-P facilitates the CBP officer's decision-making process about whether a passenger or crewmember should receive additional screening prior to entry into, or departure from, the country because that person may pose a greater risk for

⁴² See, e.g., 19 U.S.C. 482, 1431, 1433, 1461, 1496, 1499, 1581-1583; 8 U.S.C. 1221, 1357 and 49 U.S.C. 44909.

⁴³ 22 U.S.C. §5201 *et. seq.*

⁴⁴ 40 U.S.C. §1401 *et seq.*

terrorism and related crimes or other violations of U.S. law. ATS-P is a fully operational application that utilizes CBP's System Life Cycle methodology⁴⁵ and is subject to recurring systems maintenance. ATS-P is operational and has no set retirement date.

4.1.3.2. Technology and Methodology

ATS-P processes traveler information against other information available to ATS, and applies threat-based scenarios comprised of risk-based rules, to assist CBP officers in identifying individuals who require additional screening or in determining whether individuals should be allowed or denied entry into the United States. The risk-based rules are derived from discrete data elements, including criteria that pertain to specific operational/tactical objectives or local enforcement efforts. Unlike in the cargo environment, ATS-P does not use a score to determine an individual's risk level; instead, ATS-P compares information in ATS source databases against watch lists, criminal records, warrants, and patterns of suspicious activity identified through past investigations and intelligence. The results of these comparisons are either assessments of the threat-based scenario(s) that a traveler has matched, or matches against watch lists, criminal records and/or warrants. The scenarios are run against continuously updated incoming information about travelers (*e.g.*, information in passenger and crew manifests) from the data sources listed below. While the risk-based rules are initially created based on information derived from past investigations and intelligence (rather than derived through data mining), data mining queries of data in ATS and its source databases may be subsequently used by analysts to refine or further focus those rules to improve the effectiveness of their application.

The results of queries in ATS-P are designed to signal to CBP officers that further inspection of a person may be warranted, even though an individual may not have been previously associated with a law enforcement action or otherwise noted as a person of concern to law enforcement. The risk assessment analysis is generally performed in advance of a traveler's arrival in or departure from the United States, and becomes one tool available to DHS officers in determining a traveler's admissibility and in identifying illegal activity. In lieu of manual reviews of traveler information and intensive interviews with every traveler arriving in or departing from the United States, ATS-P allows CBP personnel to focus their efforts on potentially high-risk passengers. The CBP officer uses the information in ATS-P to assist in

⁴⁵ CBP's Office of Information & Technology's System Life Cycle (SLC) is a policy that lays out the documentation requirements for all CBP information technology projects, pilots and prototypes. All projects and system changes must have disciplined engineering techniques, such as defined requirements, adequate documentation, quality assurance, and senior management approvals, before moving to the next stage of the life cycle. The SLC has seven stages: initiation and authorization, project definition, system design, construction, acceptance and readiness, operations, and retirement.

determining whether an individual should undergo additional screening or should be allowed or denied entry into the United States.

4.1.3.3. Data Sources

ATS-P screening relies upon information in the DHS Advance Passenger Information System (APIS), Nonimmigrant Information System (NIIS), and Suspect and Violator Indices (SAVI); the Department of State visa databases; PNR information from commercial airlines; TECS crossing data and seizure data; and information from the consolidated and integrated terrorist watch list maintained by the FBI's Terrorist Screening Center. ATS-P uses available information from these databases to assist in the development of the risk-based rules discussed above.

4.1.3.4. Efficacy

ATS-P provides information to its users in near real time. The flexibility of ATS-P's design and cross-referencing of databases permits CBP personnel to employ information collected through multiple systems within a secure information technology system, to detect individuals requiring additional screening. The automated nature of ATS-P greatly increases the efficiency and effectiveness of the officer's otherwise manual and labor-intensive work checking individual databases, and thereby helps facilitate the more efficient movement of travelers while safeguarding the border and the security of the United States. As discussed below, ATS includes real-time updates of information from ATS source systems to ensure that CBP officers are acting upon accurate information.

4.1.3.5. Laws and Regulations

CBP is the agency responsible for collecting and reviewing information from travelers entering and departing the United States. As part of this clearance process, each traveler entering the United States must first establish his or her identity, nationality, and admissibility to the satisfaction of the CBP officer and must submit to inspection for customs purposes. The information collected is authorized pursuant to the *Enhanced Border Security and Visa Reform Act of 2002*,⁴⁶ the ATSA, the *Intelligence Reform and Terrorism Prevention Act of 2004*,⁴⁷ the *Immigration and Naturalization Act*, as amended,⁴⁸ and the *Tariff Act of 1930*, as amended.⁴⁹

⁴⁶ Pub. L. 107-173

⁴⁷ Pub. L. 108-458.

⁴⁸ 8 U.S.C § 215.

⁴⁹ 19 U.S.C. §§ 66, 1433, 1454, 1485, 1624, and 2071.

Much of the information collected can be found on routine travel documents that passengers and crewmembers currently provide to CBP when entering and departing the United States.

4.1.4. ATS Privacy Impact and Privacy Protections

The Privacy Office has worked closely with CBP to ensure that ATS satisfies the privacy documentation required for operation. CBP completed a new PIA, and published a SORN, for all six ATS modules in August 2007.⁵⁰

Authorized CBP officers, and personnel from ICE, TSA, the U.S. Citizenship and Immigration Services (USCIS), and the DHS Office of Intelligence and Analysis (DHS I&A), who are located at seaports, airports, land border ports, and operational centers around the world, use ATS to support targeting, inspection, and enforcement related requirements.⁵¹ ATS supports, but does not replace, the decision-making responsibility of CBP officers and analysts. Decisions or actions taken about individuals are not based solely upon the results of automated searches of data in the ATS system. The information obtained in such searches merely serves to assist CBP officers and analysts in either refining their analysis or formulating queries to obtain additional information upon which to base decisions or actions regarding individuals crossing U.S. borders.

ATS relies upon its source systems to ensure the accuracy and completeness of the data they provide to ATS. When a CBP officer identifies any discrepancy regarding the data, the officer will take action to correct that information, when appropriate. ATS monitors source systems for changes to the source system databases. Continuous source system updates occur in real time, or near real time, from TECS, which includes data from the National Criminal Information Center (NCIC) as well as from ACE, AMS, ACS, and AES, and APIS. When corrections are made to data in source systems, ATS updates this information immediately and uses only the latest data. In this way, ATS integrates all updated data (including accuracy updates) in as close to real time as possible.⁵²

In the event PII (such as certain data within a PNR) used by and/or maintained in ATS-P is believed by the data subject to be inaccurate, a redress process has been developed. The

⁵⁰ The PIA is available on the Privacy Office website at http://www.dhs.gov/xinfoshare/publications/editorial_0511.shtm. The SORN is also available on the Privacy Office website, at http://www.dhs.gov/xinfoshare/publications/gc_1185458955781.shtm, and in the Federal Register at 72 FR 43650 (August 6, 2007).

⁵¹ TSA, ICE, USCIS, and DHS I&A personnel have access only to a limited version of ATS.

⁵² To the extent information that is obtained from another government source (*e.g.*, Department of Motor Vehicles (DMV) data that is obtained through the National Law Enforcement Telecommunications System (NLETS)) is determined to be inaccurate, this problem would be communicated to the appropriate government source for remedial action.

individual is provided information about this process during examination at secondary inspection. CBP officers have a brochure available to each individual entering and departing the United States that provides CBP's Pledge to Travelers. This pledge gives each traveler an opportunity to speak with a passenger service representative to answer any questions about CBP procedures, requirements, policies, or complaints.⁵³ CBP has created a Customer Satisfaction Unit in its Office of Field Operations to provide redress with respect to inaccurate information collected or maintained by its electronic systems, which includes ATS. This process is available even though ATS does not form the sole basis of identification of enforcement targets.

Under the ATS SORN, CBP permits the subject of PNR or his or her representative to obtain access and request amendment of the PNR in accordance with the Privacy Act of 1974. Procedures for individuals to access ATS information are outlined in its SORN and PIA. Individuals may gain access to their own data from source systems that provide input to ATS in accordance with the procedures set out in the SORN for each source system. The FOIA provides an additional means of access to PII held in source systems. Privacy Act and FOIA requests for access to information for which ATS is the source system may be directed to CBP.

ATS underwent the C&A process in accordance with DHS and CBP policy and obtained its C&A on June 16, 2006, for a three-year period. ATS also completed a Security Risk Assessment on March 28, 2006, in compliance with FISMA, OMB policy, and National Institute of Standards and Technology (NIST) guidance.

Access to the ATS system is audited periodically to ensure that only appropriate individuals have access to the system. CBP's Office of Internal Affairs also conducts periodic reviews of the ATS system to ensure that the system is only being accessed and used in accordance with documented DHS and CBP policies. Access to the data used in ATS is restricted to persons with a clearance approved by CBP, approved access to the separate local area network, and an approved password. All CBP process owners and all system users are required to complete bi-annual training in privacy awareness and must pass an examination. If an individual does not take training, that individual loses access to all computer systems, which are integral to his or her duties as a CBP Officer. Finally, as a condition precedent to obtaining

⁵³ In addition, each traveler can visit CBP's web site (www.cbp.gov/xp/cgov/travel/customerservice/pledge_travel.xml), where specific complaints can be filed electronically. Travelers are also provided the address of the Customer Service Center where specific concerns and complaints can be addressed in writing and the telephone number of the Joint Intake Center. Travelers can also file complaints through the DHS Traveler Redress Inquiry Program (DHS TRIP) by visiting the DHS TRIP website at http://www.dhs.gov/xtrvlsec/programs/gc_1169676919316.shtm.

access to ATS, CBP employees are required to meet all privacy and security training requirements necessary to obtain access to TECS.

As discussed above, ATS both collects information directly, and derives other information from various systems. To the extent information is collected from other systems, data is retained in accordance with the record retention requirements of those systems.

The retention period for data maintained in ATS will not exceed fifteen years, after which time it will be deleted, except as noted below. The retention period for PNR, which is contained only in ATS-P, will be subject to the following further access restrictions: ATS-P users will have general access to PNR for seven years, after which time the PNR data will be moved to dormant, non-operational status. PNR data in dormant status will be retained for eight years and may be accessed only with approval of a senior DHS official designated by the Secretary of Homeland Security and only in response to an identifiable case, threat, or risk.

Notwithstanding the foregoing, information maintained only in ATS that is linked to law enforcement lookout records, CBP matches to enforcement activities, investigations or cases (i.e., specific and credible threats, and flights, individuals and routes of concern, or other defined sets of circumstances), will remain accessible for the life of the law enforcement matter to support that activity and other enforcement activities that may become related.

A National Archives and Records Administration (NARA) Electronic Records Appraisal Questionnaire was completed for Passenger Name Record (PNR) Data in spring 2005. Efforts are underway and ongoing to obtain NARA approval for the remaining data retained in ATS.

4.2. Data Analysis and Research for Trade Transparency System (DARTTS)

4.2.1. Program Description

ICE maintains the Data Analysis and Research for Trade Transparency System (DARTTS), which generates leads for and otherwise supports ICE investigations of trade-based money laundering, contraband smuggling, trade fraud, and other import-export crimes. DARTTS analyzes trade and financial data to identify statistically anomalous transactions that may warrant investigation. These anomalies are then independently confirmed and further investigated by experienced ICE investigators.

DARTTS is owned and operated by the ICE Office of Investigations Trade Transparency Unit (TTU). Trade transparency is the concept of examining U.S. and foreign trade data to identify anomalies in patterns of trade. Such anomalies can indicate trade-based money laundering or other import-export crimes that ICE is responsible for investigating, such as contraband smuggling, trafficking of counterfeit goods, misclassification of goods, and the over-

or under-valuation of goods to hide the proceeds of illegal activities. As part of the investigative process, ICE investigators and analysts must understand the relationships among importers, exporters, and the financing for a set of trade transactions, to determine which transactions are suspicious and warrant investigation. DARTTS is designed specifically to make this investigative process more efficient by automating the analysis and identification of anomalies for the investigator.

DARTTS allows ICE to perform research and analysis that is not available in any other system because of the data it contains and the level of detail at which the data can be analyzed.⁵⁴ DARTTS does not seek to predict future behavior or “profile” individuals or entities, *i.e.*, identify individuals or entities that meet a certain pattern of behavior that has been pre-determined to be suspect. Instead, it identifies trade and financial transactions that are statistically anomalous based on user-specified queries. Investigators follow up on the anomalous transactions to determine if they are in fact suspicious and warrant further investigation. Investigators gather additional facts, verify the accuracy of the DARTTS data, and use their judgment and experience in making that determination. Not all anomalies lead to formal investigations.

DARTTS is currently used only by ICE agents and analysts who work on trade, contraband, and money laundering investigations at ICE headquarters, in certain ICE field offices, and in certain attaché offices at U.S. embassies abroad. In the future, ICE plans to move DARTTS to the ICE enterprise network and ultimately to make it available to a greater number of ICE agents in the field.

4.2.2. Technology and Methodology

DARTTS uses trade data collected by other Federal agencies and foreign governments and financial data collected by CBP and the U.S. Department of the Treasury Financial Crimes Enforcement Network (FinCEN). ICE does not directly collect information from individuals or entities for inclusion in DARTTS. Instead, ICE receives data from the sources listed below via CD-ROM or external storage devices and loads the data into DARTTS. DARTTS is a stand-alone system (*i.e.*, it is not connected to any other computer systems) and does not receive any data via direct electronic transmission from another system. DARTTS data is primarily related to international commercial trade and financial transactions.

DARTTS uses COTS software to analyze raw trade and financial data to identify anomalies and other suspicious transactions. The software application is designed for

⁵⁴ For instance, DARTTS allows investigators to view totals for merchandise imports and then sort on any number of variables, such as country of origin, importer name, manufacturer name, or total value.

experienced investigators. It enables the analysis of structured and unstructured data using three tools: the drill-down technique;⁵⁵ link analysis; and charting and graphing tools that use proprietary statistical algorithms.⁵⁶ It also allows non-technical users with investigative experience to analyze large quantities of data and rapidly identify problem areas. The program makes it easier for investigators to apply their specific knowledge and expertise to ever-larger sets of data.

DARTTS performs three main types of analysis. It conducts international trade discrepancy analysis, by comparing U.S. and foreign import/export data to identify anomalies and discrepancies that warrant further investigation for potential fraud or other illegal activity. It performs unit price analysis, by analyzing trade pricing data to identify over- or under-pricing of goods, which may be an indicator of trade-based money laundering. DARTTS also performs financial data analysis, by analyzing financial reporting data (the import/export of currency, deposits of currency in financial institutions, reports of suspicious financial activities, and the identities of parties to these transactions) to identify patterns of activity that may indicate money laundering schemes.

DARTTS routinely receives bulk financial and trade information collected by other agencies and foreign governments,⁵⁷ hereafter referred to as “raw data.” The agencies that provide DARTTS with trade data collect any PII directly from individuals or enterprises completing export-import forms.⁵⁸ The agencies that provide DARTTS with financial data receive PII from individuals and institutions, such as banks, that are required to complete certain financial reporting forms.⁵⁹ The PII in the raw data is necessary, because the system uses it to

⁵⁵ The drill-down system allows investigators to quickly find, analyze, share, and document suspicious patterns in large amounts of data, and to continually observe and analyze patterns in data at any point as they progress toward the targeting goal. Investigators can also connect from one dataset within DARTTS to another, to see whether the suspicious people, entities, or patterns occur elsewhere.

⁵⁶ DARTTS provides investigators the means to represent data graphically in graphs, charts, or tables to make identification of anomalous transactions easier and visually obvious. DARTTS does not perform entity resolution nor does it create new records stored in DARTTS.

⁵⁷ Foreign Trade Data may include: names of importers, exporters, and brokers; addresses of importers and exporters; Importer IDs; Exporter IDs; Broker IDs; and Manufacturer IDs.

⁵⁸ U.S. Trade Data includes the following PII: names and addresses (home or business) of importers, exporters, brokers, and consignees; Importer and Exporter IDs (*e.g.*, an individual’s or entity’s Social Security or Tax Identification Number); Broker IDs; and Manufacturer IDs.

⁵⁹ U.S. Financial Data includes the following PII: names of individuals engaging in financial transactions that are reportable under the Bank Secrecy Act (*e.g.*, cash transactions over \$10,000); addresses; Social Security/Taxpayer Identification Numbers; passport number and country of issuance; bank account numbers; party names and addresses; and owner names and addresses.

link related transactions together. It is also necessary to identify the persons or entities that should be investigated further.

ICE investigators with experience conducting financial, money laundering, and trade fraud investigations use the completed analysis to identify possible criminal activity and provide support to field investigators. TTU investigators at ICE Headquarters refer the results of DARTTS analyses to ICE field offices as part of an investigative referral package to initiate or support a criminal investigation. ICE investigators in certain ICE field offices and attaché offices at U.S. embassies abroad also have access to DARTTS on stand-alone terminals. These investigators use DARTTS to conduct analyses in support of financial, money laundering, and trade fraud investigations, and to respond to inquiries from partner-country TTU's with whom ICE shares anonymized U.S. trade data.

4.2.3. Data Sources

All of the raw data in DARTTS is provided by other U.S. agencies and foreign governments, and is divided into three broad categories: U.S. trade data, foreign trade data, and U.S. financial data. The U.S. trade data in DARTTS is (1) import data in the form of an extract from ACS, which CBP collects from individuals and entities importing merchandise into the U.S. who complete CBP Form 7501 ("Entry Summary"); (2) export data that the U.S. Department of Commerce collects from individuals and entities exporting commodities from the U.S. using Commerce Department Form 7525-V ("Shipper's Export Declaration"); and publicly available aggregated U.S. export data (*i.e.*, data that does not include PII) purchased by ICE from the U.S. Department of Commerce.⁶⁰

The foreign import and export data in DARTTS is provided to ICE by partner countries pursuant to a Customs Mutual Assistance Agreement (CMAA) or other similar agreement. Certain countries provide trade data that has been stripped of PII. Other countries provide complete trade data, which includes any individuals' names and other identifying information that may be contained in the trade records.

ICE receives U.S. financial data from FinCEN⁶¹ for uploading into DARTTS. This data is in the form of the following financial transaction reports: Currency Monetary Instrument

⁶⁰ This dataset is further described (including a complete list of data fields) on the U.S. Commerce Department website at: <http://www.census.gov/foreign-trade/reference/products/catalog/expDVD.html>.

⁶¹ FinCEN administers the Bank Secrecy Act (BSA), a comprehensive Federal anti-money laundering statute. Pub. L. 91-508, titles I, II, October. 26, 1970, 84 Stat. 1114, 1118. The BSA requires depository institutions and other industries vulnerable to money laundering to take precautions against financial crime, including reporting financial transactions possibly indicative of money laundering.

Reports (declarations of currency or monetary instruments in excess of \$10,000 made by persons coming into or leaving the United States); Currency Transaction Reports (deposits or withdrawals of \$10,000 or more in currency into or from depository institutions); Suspicious Activity Reports (information regarding suspicious financial transactions within depository institutions and the securities and futures industry); and Reports of Cash Payments over \$10,000 Received in a Trade or Business (reports of merchandise purchased with \$10,000 or more in currency).

DARTTS itself is the source of analyses of the raw data produced using COTS software analytical tools within the system. In addition, DARTTS creates extracts of U.S. trade data that has been stripped of PII, and provides those extracts to partner countries that operate their own TTUs and with whom the United States has entered into a CMAA or other similar agreement. The U.S. financial data in DARTTS is not shared with partner countries.

4.2.4. Efficacy

The DARTTS system has proved to be a useful tool for ICE in identifying criminal activity. To date the ICE TTU has initiated several case referrals and continues to support on-going investigations. Information from the DARTTS system has assisted in several criminal prosecutions and has also identified several major money laundering schemes. For example, through information gathered with DARTTS, ICE was able to disrupt and stop a major money laundering scheme that involved the movement of several billion dollars worth of Euros from Colombia to the United States. This information resulted in two multi-million dollar seizures, which disrupted and virtually stopped the laundering of Euros into the United States from Colombia.

4.2.5. Laws and Regulations

ICE is authorized to conduct these law enforcement activities under 18 U.S.C. § 545 (Smuggling goods into the United States); 18 U.S.C. § 1956 (Laundering of Monetary Instruments); and 19 U.S.C § 1484 (Entry of Merchandise), and DHS is authorized to maintain documentation of these activities pursuant to 19 U.S.C. § 2071 note (Cargo Information) and 44 U.S.C. § 3101 (Records Management by Agency Heads; General Duties). Information in DARTTS is regulated under the Privacy Act of 1974, the Trade Secrets Act,⁶² and the Bank Secrecy Act.⁶³

⁶² 18 U.S.C. § 1905.

⁶³ Pub. L. 91-508, titles I, II, October 26, 1970, 84 Stat. 1114, 1118.

4.2.6. Privacy Impact and Privacy Protections

ICE does not use DARTTS to make unevaluated automated decisions about individuals, and DARTTS data are never used directly as evidence to prosecute crimes. DARTTS is solely an analytical tool that helps in the identification of anomalies. It is incumbent upon the investigator who finds an anomaly to further investigate the reason for the anomaly. If the anomaly can be legitimately explained, the investigator has no need to further investigate it for criminal violations and moves on to the next identifiable anomaly. In addition, ICE investigators are required to obtain and verify the original source data from the agency that collected the information, to prevent inaccurate information from propagating. All information obtained from DARTTS is independently verified before it is acted upon or included in an ICE investigative or analytical report. Investigators follow up on anomalous transactions to determine if they are in fact suspicious and warrant further investigation. They gather additional facts, verify the accuracy of the DARTTS data, and use their judgment and experience in making that determination.

DARTTS data generally is subject to access and amendment requests under the Privacy Act of 1974 and the FOIA, unless a statutory exemption covering specific data applies. The U.S. and foreign government agencies that collect the information uploaded into DARTTS are responsible for providing appropriate notice on the forms used to collect the information and/or through other forms of public notice, such as SORNs.⁶⁴ DARTTS will coordinate requests for access or to amend data with the original data owner. ICE has worked closely with the Privacy Office to complete and publish a PIA and a SORN for DARTTS.⁶⁵

As all of the information in DARTTS is obtained from other governmental organizations that collect the data under specific legislative authority, DARTTS cannot independently verify the accuracy of the data it receives. The owner of the source data is responsible for maintaining and checking the accuracy of its own data. In many instances, the data ultimately loaded into

⁶⁴ The following SORNs are published in the Federal Register and describe the raw data ICE receives from U.S. agencies for use in DARTTS: for FinCEN Information, Suspicious Activity Report System (Treasury/FinCEN .002) and Bank Secrecy Act Reports System (Treasury/FinCEN .003); for Commerce Department Information, Individuals Identified in Export Transactions System (Commerce/ITA-1); and for CBP Information, Automated Commercial Environment/International Trade Data System (ACE/ITDS) (DHS/CBP-001).

⁶⁵ The PIA is available on the Privacy Office website at http://www.dhs.gov/xinfoshare/publications/editorial_0511.shtm#6. DARTTS is covered by the SORN for the ICE Trade Transparency and Analysis Research (TTAR) system of records. The SORN is available on the Privacy Office website at http://www.dhs.gov/xinfoshare/publications/gc_1185458955781.shtm#5, and in the Federal Register at 73 FR 64967 (October 31, 2008).

DARTTS is highly accurate because it is collected directly from the individual. In other instances, however, the data about individuals is provided to a governmental organization by a third party. In the event that errors are found, the DARTTS system owner must notify the agency that originally collected the data. FinCEN currently provides ICE with corrections to existing data, which are then uploaded into DARTTS. ICE does not, however, receive data corrections on trade data.

DARTTS received its C&A from DHS IT Security on September 28, 2006, and the C&A is in effect for three years. DARTTS is maintained in a secure, government-owned facility. It is a stand-alone system, *i.e.*, it is not networked to any other internal or external computer network. This provides a high degree of security against unauthorized access through hacking or other means. Any violations of systems security or suspected criminal activity will be reported to the DHS Office of Inspector General, the Office of the Information System Security Manager (OISSM) team in accordance with the DHS security standards, and to the ICE Office of Professional Responsibility.

All DARTTS users are assigned unique user IDs and passwords. Audit trails are used to track user activities and provide accountability. Only authorized personnel can access audit trails, which are kept for a minimum of 90 days. Audit trails are reviewed by DARTTS system administrators or the Information System Security Officer (ISSO). The system administrator also maintains a spreadsheet record of the receipt or distribution of sensitive information on electronic media.⁶⁶

Access to DARTTS is granted on a case-by-case basis by the TTU Network Administrator. Access is currently limited to ICE Special Agents and Criminal Research Specialists who work on TTU investigations at ICE Headquarters or in the ICE field and foreign attaché offices, as well as properly cleared support personnel. Access is further limited only to individuals who have physical access to the offices of the TTU at ICE Headquarters, or access to the ICE field offices and foreign attaché offices that have stand-alone DARTTS terminals. ICE is exploring the feasibility of placing DARTTS on the ICE enterprise network in order to further expand access to other ICE agents in the field.

ICE is in the process of drafting a proposed record retention schedule for the information maintained in DARTTS. ICE anticipates maintaining the records in DARTTS for five (5) years and then archiving records for five additional years, for a total retention period of ten (10) years.

⁶⁶ DARTTS receives CD-ROMs and other external storage media provided by other agencies in a password-protected encrypted format. Once data from CD-ROMs or other external storage media is loaded onto DARTTS, the TTU Network Administrator stores them in the secured server room located in the TTU offices at ICE Headquarters until the retention period has elapsed, at which point they are destroyed.

The five-year retention period for records is necessary to create a data set large enough to effectively identify anomalies and patterns of behavior in trade transactions. Records older than five (5) years will be removed from the system and archived for five (5) additional years and will only be used to provide a historical basis for anomalies in current trade activity. The original CD-ROMs containing the raw data will be retained for five (5) years to ensure data integrity and for system maintenance.

4.3. Freight Assessment System (FAS)

4.3.1. Program Description

The TSA Freight Assessment System (FAS) is a risk-assessment tool that can be used to identify cargo that may pose a heightened risk to passenger aircraft. To reduce the current reliance on random inspections, FAS uses a rules-based model developed by security subject-matter experts that is software-based and incorporates machine-derived rules and predictive indicators to identify and assess high-risk cargo. Once FAS is fully operational, cargo identified as high-risk will be flagged and set aside for further inspection by air carriers. The FAS system neither uses nor stores PII.

FAS was originally designed to support the mandates of the 2003 TSA Air Cargo Strategic Plan. Section 1602 of the *9/11* Commission Act, which was enacted during the FAS design and development process, added new requirements to the Air Cargo strategic goals and mission, specifically, that by February 2009, 50 percent of all cargo destined for passenger aircraft must be screened, and 100 percent by August 2010.⁶⁷ The FAS will be another layer of security, because it identifies domestic cargo for secondary screening and assesses risk for international inbound cargo as well. FAS facilitates government-managed, risk-based assessments of cargo and chain-of-custody oversight throughout the supply chain, including locations far from the air carrier where the risk is more significant.

With help from selected industry participants in the air cargo supply chain, FAS has completed the pre-system testing described in the 2007 Report. The results indicate that implementing FAS in participants' business operations has a minor impact on the efficiency of those operations. FAS obtained its C&A from DHS/TSA IT Security on May 23, 2008, and has received its Authority to Operate (ATO) for 2 years. FAS's development stage is now concluded, and FAS is currently being prioritized for operational implementation. The planned FAS life cycle is 10 years.

⁶⁷ To fulfill these requirements, the Department will publish regulations to mandate the screening of all cargo carried on passenger aircraft.

4.3.2. Technology and Methodology

FAS uses COTS software to collect, fuse, and analyze data related to supply chain, logistics, and freight transportation data. The software serves as the information platform that supports TSA's layered security approach. It (1) applies risk-based analytical rules to the supply chain data to assist analysts in scrutinizing cargo as it is being screened, (2) provides an informational backbone that supports chain-of-custody integrity, and (3) provides business intelligence to assist management in determining the optimal usage of inspection resources. The software includes a rules-management platform that enables the application of business process, security, and data cleansing rules to each shipment. It also includes a process for incorporating information from screeners and analysts, based upon their real-world experience, into the risk-based rules development process.

4.3.3. Data Sources

FAS uses the air carriers' house and master airway bills (no PII from airway bills is included in the system), and data from the following TSA systems: the Performance and Results Information System (PARIS) (compliance data);⁶⁸ the Indirect Air Carrier Management System (IACMS) (TSA-assigned certification number);⁶⁹ and the Known Shipper Management System (KSMS) (company names).⁷⁰ FAS compares information in these TSA systems with company background information that it obtains from Dun & Bradstreet, and with publicly-available statistical data on criminal activity.

⁶⁸ PARIS compiles the results of cargo inspections and the actions taken when violations are identified. The PARIS database provides TSA a web-based method for entering, storing, and retrieving performance activities and information on TSA-regulated entities, including air carriers and indirect air carriers. PARIS includes profiles for each entity, inspections conducted by TSA, incidents that occur throughout the nation, such as instances of bomb threats, and investigations that are prompted by incidents or inspection findings.

⁶⁹ IACMS is a management system used by TSA to approve and validate new and existing Indirect Air Carriers. This management system and application is intended for freight forwarders wishing to receive TSA approval to tender cargo utilizing an Indirect Air Carrier certification. The IACMS is not intended for individuals wanting to ship cargo. An Indirect Air Carrier means any person or entity within the United States not in possession of a Federal Aviation Administration air carrier operating certificate that undertakes to engage indirectly in air transportation of property and uses, for all or any part of such transportation, the services of a passenger air carrier. See PIA on TSA's Air Cargo Security Requirements, published on the DHS Privacy Office web site on May 25, 2006, which provides additional information on the privacy impact of the IACMS and KSMS. The PIA is available on the Privacy Office website at http://www.dhs.gov/xinfoshare/publications/editorial_0511.shtm.

⁷⁰ KSMS uses commercial databases to verify the legitimacy of shippers. Known shippers are entities that have routine business dealings with freight forwarders or air carriers and are considered vetted shippers. In contrast, unknown shippers are entities that have conducted limited or no prior business with a freight forwarder or air carrier.

4.3.4. Efficacy

The software used in FAS has proven in various testing and operational deployments to be highly effective in providing accurate information in real time that supports TSA risk analysts (the primary users of the system) in their work. The value of the output generated by the software will be determined by the quality of the data inputs, including airway bills, the primary document used in FAS. Where necessary, FAS can use the software's industry-specific capabilities (supply chain, logistics, and freight transportation) to enable data cleansing and the interpretation and improvement of any data that appears to be of lesser quality.

Prior to the testing phase described in the 2007 Report, validity testing was completed in a proof of concept that scored live airway bills with a prototype model of FAS. TSA found the results to be consistent with expectations and with CBP's ATS determinations. TSA was able to inspect cargo that the tools identified as presenting an elevated risk, and physical inspection confirmed the higher risk determination. Standards for validating the data mining models are included in the software used by FAS. The results of the FAS testing phase corroborate the results of the proof-of-concept validity testing.

4.3.5. Laws and Regulations

The legal and policy foundation for FAS is based in the ATSA, which established TSA and gave it responsibility for security in all modes of transportation;⁷¹ the DHS/TSA Air Cargo Strategic Plan (November 2003), which sets forth TSA's commitment to work closely with Federal, state, local, and industry partners to ensure that 100 percent of cargo that is deemed to be of elevated risk is inspected and that 100 percent of the cargo supply chain is secure; and Section 1602 of the 9/11 Commission Act, which requires TSA to provide a level of security for cargo on commercial flights that is commensurate with the level of security provided for passenger checked baggage. The 9/11 Commission Act also sets the inspection benchmarks of 50% of cargo screened not later than 18 months after the date of enactment, and 100% of cargo screened not later than three years after date of enactment.

4.3.6. Privacy Impact and Privacy Protections

FAS has completed a PTA in conjunction with the Privacy Office. As TSA specifically designed the FAS system not to hold any PII, individual privacy is not affected by FAS and no privacy documentation beyond the PTA is required. Dun & Bradstreet data is used to confirm that the Indirect Air Carrier certification issued to a business owner matches the name on the certificate issued by TSA's Indirect Air Carrier regional coordinators. The information regarding

⁷¹ Pub. L. 107-71 (2001).

the business owner is not entered into FAS as a factor. It is only used to verify the name of the business owner.

FAS has obtained its C&A from DHS/ TSA IT Security, demonstrating its full compliance with DHS information technology security requirements. Access to the FAS system is limited to authorized users. Unauthorized access is controlled through system lockdowns, role-based access control, and the use of authentication servers. HTTPS (TLS) will be the only acceptable method of communication with the web server. Audit capability exists within FAS and will be used to review the reasons for system hits.

As noted above, FAS is currently being prioritized for operational implementation. FAS implementation plans call for retention of data in FAS for a period of 90 days for examination and analysis, after which the data will be archived for 7 years.

5. Public Workshop on Privacy Protections for Government Data Mining

In the 2007 Report, the Privacy Office stated its interest in holding a public workshop to explore the privacy issues associated with government data mining, and to inform the Office and the public about available technological means, such as anonymization and automated audit tools, of addressing those issues. The workshop, *Implementing Privacy Protections in Government Data Mining*, took place on July 24-25, 2008.⁷² The discussion that follows summarizes the views of the workshop participants.

Panelists included academics, government researchers, policy and technology experts, and privacy advocates. Speakers described the actual and potential impacts of government data mining on individuals and on society. They explored methods of validating the accuracy and effectiveness of data mining models and results, and they discussed anonymization tools and audit controls that could be implemented to enhance privacy protection in data mining projects. Finally, they identified key elements of best practices for engaging in data mining activities in a manner that both protects privacy and furthers the Department's mission.

⁷² The Federal Register Notice announcing the workshop, the workshop agenda and transcript, and public comments submitted in response to the Notice are all available on the Office's website at http://www.dhs.gov/xinfoshare/committees/editorial_0699.shtm.

The panelists identified several key lessons for government data mining:

- Data mining technology can be an important tool for protecting national security. Where activities involving data mining are effective and designed in a manner that limits impacts on individual privacy, they should be conducted to further the Department's mission.
- There is a need for clear rules for designing data mining programs and establishing their effectiveness, for determining how the results of data mining are used, for monitoring the privacy impacts of projects involving data mining, and for providing redress where appropriate.
- As the consequences for individuals falsely identified as possible criminals or terrorists can be devastating, the number of such "false positives" in the results of data mining must be minimized.
- Validating data mining models and results is essential to the effectiveness of data mining programs. The validation should evaluate not only the algorithms used, but also all aspects of a data activity – from data collection through decision-making based upon the results.
- Activities that use data mining should be transparent and should include both automated and non-automated monitoring for compliance with legal rules and internal controls.
- Objective oversight of government data mining activities is essential to ensure their integrity and efficacy. Oversight could take the form of internal reviews (*e.g.*, by government employees other than the activities' proponents), reviews by entities such as institutional review boards or expert advisory committees, or legislation.
- The results of data mining should be used for investigative purposes only, and not as the sole basis for decision-making about individuals. Data mining can be a powerful analytical tool, but it is not a substitute for the exercise of judgment by professional investigators.

5.1. Defining "Data Mining"

The workshop began with a presentation on defining data mining and on the policy issues that flow from that definition. Data mining uses mathematical algorithms to construct statistical models that estimate the value of an unobserved variable-- for example, the probability that an individual will engage in illegal activity. Data mining is best understood as an iterative process consisting of two separate stages: machine learning, where algorithms are applied against known

data; and probabilistic inference, where the models built from algorithms are applied against unknown data to make predictions.

The more advanced data mining applications thus yield estimates of probability rather than binary yes/no classifications. In the case of data mining to uncover illegal activity, probabilistic inference makes it possible to focus on those targets identified as having a higher likelihood of association with illegal activity. The inference process is conducted in successive stages, thereby yielding an increasingly sharper focus on high-probability targets and improving accuracy. The utility of data mining models should be assessed by comparing the results to the results of rules-based decisions made by investigators in the field.

Current definitions of data mining, such as the Data Mining Reporting Act's distinction between "pattern-based" electronic searches (which are deemed "data mining") and "subject-based" searches (which are not), are too narrow. The multi-stage inference process in data mining can incorporate both "pattern-based" and "subject-based" searches. Moreover, privacy issues are inherent regardless of the type of search performed. For the purpose of addressing the privacy and public policy issues, it is important to define data mining broadly, to include the institutional context in which the mathematical tools and research models are used, from data-gathering and inferences drawn about potential targets, to decision-making about individuals.

5.2. Panel on How Government Data Mining Impacts Privacy

The panel on privacy impacts focused on the potential negative effects of data mining on individuals and on society at large, where data mining is conducted without safeguards. The potential harms are related to data quality issues and to imprecision in data mining methodology. The use of inaccurate data for data mining can lead to the misidentification of individuals as subjects of interest to law enforcement. An equally serious concern is the likelihood of "false positive" results (*i.e.*, results stemming from an inaccurate data mining model that falsely identifies individuals as potential criminals or terrorists). The consequences of such misidentification for individuals can range from mere inconvenience to much more devastating outcomes, up to and including incarceration.

Panelists identified several potential societal harms that could be caused by government data mining that is not conducted with appropriate safeguards. Significant public resources would be wasted where, for whatever reason, data mining misidentified individuals as potential criminals or terrorists, leading law enforcement personnel to focus on the wrong targets, and leaving actual threats unaddressed as a result. There are also political risks: potentially valuable data mining programs undertaken without the necessary protections lose public support and funding; and talented investigators and researchers are discouraged from designing valuable programs when they are denied funding in the absence of clear rules for what is acceptable and what is not.

Panelists also identified potential civil liberties harms from data mining that lacks appropriate safeguards. Data mining models that inaccurately associate potential criminal or terrorist activity with racial or ethnic characteristics would stigmatize entire classes of individuals. As well, individuals could adjust their behavior in ways inconsistent with a free society, for example, by refraining from associating with persons from certain ethnic groups, to avoid the consequences of government data mining erroneously based on ethnic or racial characteristics.

5.3. Panel on Validating Data Mining Models and Results

Speakers on the panel on validating data mining research were engineers and computer scientists from the DHS Science and Technology Directorate and from the private sector. They discussed how the complex process of validating data mining models should be carried out. First, panelists stated, data mining algorithms used by the government should be made publicly available, even as the data these algorithms act on, and the conclusions reached based upon them, remain non-public. According to several panelists, following the “open-source” paradigm would subject data mining algorithms to the salutary effects of trouble-shooting by external technology experts.

The type of data used to test data mining models has been a significant issue. While some privacy advocates argue that research should be carried out using synthetic -- *i.e.*, simulated-- data, the panelists agreed that synthetic data is generally a very simplified model of reality and thus useful only in the earliest stages of model-building (*e.g.*, to see if an algorithm or model works at all, or to test for scalability). Testing a model with synthetic data gives very little assurance that the model will be effective in assessing real-life situations. Therefore, according to several panelists, real data is necessary to model actual conditions and to improve an algorithm’s performance.

The statistical methods of testing the accuracy of data mining models are well established. “Supervised” models are built by running algorithms against large data sets that include information about previously known behaviors, and then testing them against different data sets to see how well they predict those behaviors. “Unsupervised” data mining models are used when information about a particular behavior is incomplete or lacking, as is currently the case for acts of terrorism. To build an unsupervised model, it is necessary first to test the efficacy of an existing predictive or investigative process (*e.g.*, by ascertaining its false positive rate) and then to measure the model’s efficacy against it. The model’s predictive power must be measured against actual human experience. This would be particularly true for models designed to predict the likelihood that an individual or group of individuals is likely to be involved in criminal or terrorist activity.

Panelists argued for an expansive validation process. They stated that validation of data mining models and results must not be limited to the algorithms themselves; the larger context in which algorithms are used --from data gathering to decision-making-- must also be scrutinized. Data mining is best thought of as a “power tool” for analysis, to be used in conjunction with human judgment based on experience. It should be used to minimize the number of people identified for further investigation, not to flag people solely on the basis of the results. Finally, as data mining is an iterative process, it can be self-correcting to some extent: while it is not possible to totally eliminate false positives resulting from an initial pass through the data, subsequent passes can minimize the number of false positives.

Panelists noted that the quality of the data used for data mining is central to the integrity of the outputs; thus, validation begins with careful preparation and ongoing review of the data set. In addition, the model’s effectiveness must be reviewed by comparing it to the results of human decision-making. Finally, researchers must address system-level concerns such as usability, auditability, and information assurance (authentication of users, role-based permissions) in preparation for conducting data mining research.

5.4. Panel on Technologies for Privacy-Protective Data Mining

The panel on privacy-protective technologies surveyed the current research and development status of automated tools that could be used to enhance the privacy of information used in data mining activities. Panelists discussed techniques that enable sophisticated analyses to be performed on large data sets without access to PII, either because that information has been stripped of identifiers (anonymization) or because it has been obscured by adding “noise” to the data or by other means (randomization or data transformation). Panelists also discussed “secure multi-party computation,” whereby “owners” of distributed databases can collaborate in analyses of the distributed data, using secure protocols, without actually collecting each other’s data. This method would enable searches for anomalies or commonalities in data in distributed databases, where only the results of the analyses are shared. The protocols could use cryptographic standards, randomization, or anonymization.

Some automated anonymization techniques show promise for minimizing the use of PII in data mining, but more research is needed to determine whether they are, in fact, capable of being implemented in a privacy-protective manner. Deployment has been slow, in part due to the complexity of the software applications themselves, and in part due to the challenges of running the software concurrently with complex data mining algorithms. Panelists identified several goals for further research, including (1) designing anonymization techniques that cannot be reversed; (2) reducing the functional/technical impacts of the technology on the efficiency of data mining algorithms and models; and (3) integrating privacy policy into software.

Finally, panelists agreed on the desirability of a comprehensive approach to addressing the challenge of building privacy protections for data mining. Data mining technology itself is not the only source of potential risks to privacy, and so research is needed to design systems to build privacy into the entire life cycle of personal information used in data mining activities. Collection, data preparation and storage, the use of personal information in data mining analyses, and, ultimately, the decisions about individuals that may be based upon the results, should all be carried out in a manner that protects privacy.

5.5. Panel on Auditing and Enforcing Privacy Controls for Data Mining

Speakers on the panel on auditing and enforcing privacy controls agreed, first, that data mining is important for national security purposes, and, second, that compliance-monitoring mechanisms must be an integral part of data mining activities to ensure that they are conducted in a manner that satisfies legal and policy controls. The core elements of audits traditionally performed by accounting firms-- *e.g.*, reviews of the efficacy of internal controls applicable to personal information, and assessments of compliance with legal, regulatory, and policy standards -- present a good model for monitoring a data mining activity's compliance with privacy rules. Automated audit and compliance-monitoring systems are increasingly necessary, in light of the large-scale systems and databases currently used in data mining and the complexity of the data mining activities themselves. Automated systems should serve as a supplement to, rather than take the place of, human judgments about compliance.

Academic researchers on the panel presented a prototype "proof-of-concept" of software that integrates both access-control techniques and legal or policy rules into large-scale database systems. The software logs transactions -- *i.e.*, individuals accessing the data, and their use of the data -- and runs machine-readable rules or policy governing those actions in tandem with the system logs. It flags those access or usage "events" that, when compared with the rules, indicate potential violations, and provides information explaining why a violation may have occurred. The software thus assists in detecting anomalies meriting further investigation through non-automated means.

Panelists agreed that for any audit or compliance-monitoring mechanism to be effective, the privacy rules and standards against which data mining activities are measured must be consistent and clearly articulated (and, for automated systems, capable of being translated into machine-readable form). As well, security measures in place to protect systems used in data mining must be continuously monitored and assessed for effectiveness.

5.6. Panel on Privacy Policies and Best Practices for Government Data Mining

Speakers on the panel on privacy policies and best practices discussed privacy standards for government data mining. Building on the current consensus among many government and

private-sector groups, academics, privacy advocates, security experts, and others who have studied the privacy issues associated with data mining, the panelists identified the following privacy requirements for conducting data mining programs:

- Authorization by Congress or a senior administration official.
- Maximum transparency consistent with program objectives.
- Collection and use of data consistent with applicable law.
- Assessment of program effectiveness (including impacts on individuals) in achieving documented purpose(s), both before and during deployment.
- Limitations on access by government employees or contractors to databases used for data mining.
- Limitations on the purposes for which the data is used, to prevent “mission creep.” The need for purpose limitation is particularly acute where the government obtains data originally collected by the private sector.
- Use of data minimization and anonymization tools, to limit the information accessed in data mining programs.
- Audit tools to monitor compliance with rules and policies governing data mining and the data used.
- Review and approval of data mining programs conducted by individuals other than the proponents of those programs, *e.g.*, by an institutional review board or an external advisory committee.
- Redress mechanisms to address harms caused where data mining results are erroneously applied to individuals. Redress could include a feedback mechanism, similar to that in the credit-reporting context, which allows for continuous correction of data in databases used for data mining.
- Rigorous oversight of data mining programs. Oversight could take various forms, including review by agency privacy officers or by agency inspectors general, as well as through legislation.

As discussed below, the Department has incorporated these protections into a process for reviewing privacy-sensitive research projects conducted by its Science and Technology Directorate.

6. Privacy Principles for DHS S&T Research Projects

6.1. Background

The Privacy Office has been working with the DHS Science and Technology Directorate (S&T), the Department's primary research and development arm, to identify areas of mutual interest and the potential for collaboration in support of the Department's mission. S&T organizes the scientific and technological resources of the United States to prevent or mitigate the effects of catastrophic terrorism against the United States or its allies. It both conducts its own research and works in partnership with the private sector and other government agencies to encourage innovation in homeland security research and technology development. S&T research encompasses a wide range of activities, from biological research on animal diseases, to social-behavioral research on the motivations of terrorism, to the development of new physical screening technologies. Many S&T research projects do not involve or impact PII; however, when PII is involved, S&T works closely with the Privacy Office to ensure that all privacy-sensitive S&T research projects safeguard PII and protect the privacy of individuals. S&T worked closely with the Privacy Office to develop the agenda for the Data Mining Workshop, and S&T has been very supportive of the Privacy Office's goal of developing data mining guidance for the Department.

As an outgrowth of the Data Mining Workshop, the Privacy Office and S&T began working to identify privacy principles that could be embedded in S&T's research and development projects involving data mining, to ensure that those projects would be carried out in a manner that both protects privacy and furthers the Department's mission. As this effort progressed, S&T and the Privacy Office determined to broaden the analysis to include to all privacy-sensitive S&T research. This collaboration has led to a set of *Principles for Implementing Privacy Protections in S&T Research* (Principles), which S&T has agreed will govern new research performed at S&T laboratories, S&T-sponsored research conducted in cooperation with other Federal government entities, and research conducted by external performers under a contract with S&T.⁷³

6.2. Privacy Principles for DHS S&T Research Projects

Consistent with privacy implementation generally at the Department, the Principles reflect the Fair Information Practice Principles of Transparency, Individual Participation, Purpose Specification, Data Minimization, Use Limitation, Data Quality and Integrity, Security, and Accountability and Auditing. The Principles developed for S&T research include an

⁷³ The full text of the Principles is included in the Appendix.

additional principle, the *Privacy Assessment Principle*, which reinforces that an assessment of privacy impacts, conducted jointly by S&T and the Privacy Office, will be an integral part of the development and implementation of any S&T research project that is privacy-sensitive and/or involves or impacts PII. The Principles support taking privacy impacts into account from a project's inception, beginning with the design phase. The *Purpose Specification Principle* states that a project's purpose will be clearly articulated, and documented, through a process that includes reviews of the project's effectiveness by experts from within and outside the Department. A key premise is that the project's research activities must be within the scope of its articulated purpose. In keeping with the Department's commitment to informing the public about its information practices, the *Transparency Principle* states that S&T will conduct PIAs, in conjunction with the Privacy Office, for all research projects that involve or impact PII, and will publish PIAs for all non-classified research.⁷⁴

A primary goal of the Principles is to preclude the possibility that research projects could have a negative impact on privacy. This requires not only a tightly-focused purpose for research, but also reasonable limits on the types of data used, and on how the data is used, consistent with a project's purpose. Thus, the *Data Quality and Integrity Principle* limits a project's use of data to that which is reasonably considered both accurate and appropriate for the project's documented purpose(s), and the *Data Minimization Principle* requires that projects use the least amount of PII consistent with their documented purpose(s). The Principles also support the use of data minimization techniques to accomplish this goal, where practicable. Similarly, the *Use Limitation Principle* provides that projects will only use data in a manner that is consistent with disclosures in all applicable PIAs and SORNs, and consistent with privacy notices and policies that apply to data originally collected by the private sector.

The credibility and effectiveness of S&T research projects depends in large measure on public confidence that the data involved will be protected from unauthorized use or disclosure. To that end, the *Data Security Principle* requires researchers to take all reasonable steps necessary to maintain the security of the data they use. It is also necessary that there be public confidence that applicable rules and policies for use of that data are followed. The Principles address this concern by calling for (1) training for project personnel on Department privacy policy generally, and on the privacy protections built into research projects as a result of the PIA process (*Training Principle*), and (2) the use of automated and/or non-automated audit procedures, as appropriate, to ensure compliance with project access and data usage rules (*Audit Principle*).

⁷⁴ The Privacy Office approves PIAs for classified research, but such PIAs are not published.

Finally, an essential key to public trust in S&T's research efforts is a public understanding that individuals who believe they have been adversely affected by the research have a way to raise their concerns and to obtain relief where warranted. A redress program not only serves this purpose but can also provide valuable feedback to project researchers about problems that should be addressed systemically. Therefore, the Privacy Office, in conjunction with S&T's Privacy Officer, will develop and administer a redress program to handle inquiries and complaints regarding any S&T research project (*Redress Principle*).

The Privacy Office and S&T will work together to create an implementation plan setting forth general guidance regarding the application of these Principles to new S&T Projects.⁷⁵ In addition, the Privacy Office will continue its current practice of assessing each S&T Project through a PTA. The PTA provides a mechanism for determining whether a research project is privacy-sensitive and/or involves or impacts PII, and whether a PIA will be required. During the PTA review process, the Privacy Office and S&T will jointly determine and document how best to apply the Principles to each S&T Project.

7. Conclusion

The Privacy Office is pleased to provide Congress its third report on DHS data mining activities, and its first comprehensive report pursuant to the Data Mining Reporting Act. To ensure that the Department's use of technology sustains and does not erode privacy, as Section 222 of the Homeland Security Act mandates, the Privacy Office requires all DHS programs, including those that engage in data mining, to conduct a Privacy Threshold Analysis. Where the Privacy Threshold Analysis determines that a program is privacy-sensitive and/or involves or impacts PII, the Privacy Office works closely with the relevant component to complete a Privacy Impact Assessment that identifies the privacy impacts of the technology being used and explains the measures taken to mitigate those impacts. The Privacy Office has used these analytical tools to assess and approve the steps taken by each of the programs described in this report to protect privacy. The Privacy Office will continue to provide vigilant oversight in this area, as it does for all DHS programs.

As the Privacy Office's Data Mining Workshop demonstrated, the term "data mining" can mean different things to different people. One thing is clear, however: regardless of how data

⁷⁵ On October 7, 2008, the National Research Council published a report entitled *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment* ("NRC Report"). The NRC Report provides an analytical framework for assessing the effectiveness, legal compliance and consistency with U.S. values of government data mining and behavioral surveillance programs. The Principles that the Privacy office has developed with DHS S&T are substantially similar to the elements of the NRC Report framework, and the Privacy Office will take the framework elements into account as it works with S&T to build the plan for implementing the Principles.

mining is defined, data mining research that uses PII can have significant impacts on individual privacy, and those impacts must be addressed. The Department has taken a major step toward this goal by developing its *Principles for Implementing Privacy Protections for Research Projects*, which will be embedded in new research projects carried out by S&T, whether they involve data mining or not. The Privacy Office looks forward to collaborating with S&T to implement these Principles, so that research critical to the Department's mission is carried out in a manner that sustains individual privacy.

8. Appendix

Principles for Implementing Privacy Protections in Research Projects Science and Technology Directorate United States Department of Homeland Security

INTRODUCTION

The Department of Homeland Security's (DHS) Privacy Office and Directorate of Science and Technology (S&T) have developed these Principles to provide a privacy-protective framework for conducting critical homeland security research and development. The Privacy Office operates under the direction of the Chief Privacy Officer, who is appointed by and reports directly to the Secretary of the Department. The Office serves to implement Section 222 of the *Homeland Security Act of 2002*,¹ and has programmatic responsibilities involving the *Privacy Act of 1974*,² the *Freedom of Information Act* ("FOIA"),³ the privacy provisions of the *E-Government Act of 2002*,⁴ and DHS policies that protect individual privacy associated with the collection, use, and disclosure of PII.⁵ Section 222 of the *Homeland Security Act of 2002* calls on the Chief Privacy Officer to assume primary responsibility for privacy policy within the Department, and, among other things, to "[assure] that the use of technologies sustain, and do not erode, privacy protections relating to the use, collection, and disclosure of personal information."⁶

S&T is the Department's primary research and development arm. S&T organizes the scientific and technological resources of the United States to prevent or mitigate the effects of catastrophic terrorism against the United States or its allies. It both conducts its own research and works in partnership with the private sector and other government agencies to encourage innovation in homeland security research and technology development. S&T research encompasses a wide range of activities, from biological research on animal diseases, to social-behavioral research on the motivations for terrorism, to the development of new physical screening technologies. Many

¹ Section 222 of the Homeland Security Act of 2002, as amended by Section 8305 of the Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. 108-458 (December 17, 2004), 6 U.S.C. § 142.

² 5 U.S.C. § 552a.

³ 5 U.S.C. § 552.

⁴ Pub. L. 107-347, 116 STAT. 2899, § 208; 44 U.S.C. Chapter 36.

⁵ "Personally identifiable information" is defined as any information that permits the identity of an individual to be directly or indirectly inferred, including any information which is linked or linkable to that individual regardless of whether the individual is a U.S. citizen, a Legal Permanent Resident, or a visitor to the U.S.

⁶ 6 U.S.C. § 142(a) (1).

S&T research projects do not involve or impact PII; however, when PII is involved, S&T works closely with the Privacy Office to ensure that all privacy-sensitive S&T Projects safeguard PII and protect the privacy of individuals.

The Principles set forth below are intended to ensure that S&T builds privacy protections into the design and implementation of its research and development projects in a manner that supports the Department's mission. These principles govern research performed at S&T laboratories, S&T-sponsored research conducted in cooperation with other Federal government entities, and research conducted by external performers under a contract with S&T (collectively referred to as "Projects"). These principles apply specifically to privacy-sensitive projects.

The Privacy Office and S&T will work together to create an implementation plan setting forth general guidance regarding the application of these Principles to new S&T Projects. In addition, the Privacy Office will continue its current practice of assessing each S&T Project through a Privacy Threshold Analysis (PTA). The PTA provides a mechanism for determining whether a given Project is privacy-sensitive, and/or involves or impacts personally identifiable information (PII), and whether a Privacy Impact Assessment (PIA) will be required under the *E-Government Act of 2002*. During the PTA process, the Privacy Office and S&T will jointly determine and document how best to apply these Principles to each S&T Project.

PRINCIPLES

- Privacy Assessment Principle: An assessment of privacy impacts will be integral to the development and implementation of any research project.
 - The Privacy Office will assist S&T in identifying privacy impacts to address in project design and implementation, to ensure that research projects sustain privacy protections relating to the use, collection, and disclosure of PII pursuant to Section 222(a)(1) of the Homeland Security Act. An appropriately cleared S&T or external expert will participate in the privacy assessment to explain scientific aspects of a proposed research project where a deeper understanding is needed to make decisions regarding the use of PII.
 - All projects will complete a Privacy Threshold Analysis (PTA). Privacy Office staff and the S&T Privacy Officer will review each PTA to determine how best to apply these Principles to each project.
- Purpose Specification Principle: A project's purpose will be clearly articulated and documented through an internal/external project review process.
 - Legal Authorization: Projects will be structured to function consistent with all relevant legal requirements.
 - Purpose Limitation: Projects will only engage in research that is within the scope of their documented purpose(s).
 - Effectiveness Reviews: Projects determined by the Chief Privacy Officer to be privacy-sensitive will include an initial review before commencement of research involving PII. The review will be both internal (by S&T staff other than the project's proponents) and external (by experts with appropriate security clearances), and will assess the project's likely effectiveness in accomplishing the documented purpose(s).
- Data Quality and Integrity Principle: Projects will endeavor to only use PII that is reasonably considered accurate and appropriate for their documented purpose(s), and to protect the integrity of the data.
 - Projects will exercise due diligence in evaluating the accuracy and relevance of any publicly-available or commercially-available data used, to ensure the research effort's soundness and the integrity of the research results.

- Data Minimization Principle: Projects will use the least amount of PII consistent with their documented purpose(s), and will use PII minimization techniques such as synthetic data or anonymization where appropriate and practicable.
- Use Limitation Principle: Projects will use PII consistent with all applicable System of Records Notices (SORNs), Privacy Impact Assessments (PIAs), and other privacy notices and policies, regardless of the source of the data (*i.e.*, whether the data is collected directly by the Department of Homeland Security or its contractors, or is obtained by the Department or its contractors from third-party sources).
- Data Security Principle: Projects will take all reasonable steps necessary to maintain the security of the PII they use, and to protect the data from inappropriate, unauthorized, or unlawful access, use, disclosure, or destruction.
- Audit Principle: Projects involving PII will employ automated and/or non-automated auditing procedures, as appropriate, to ensure compliance with project access and data usage rules (“rules” are specific instructions implementing an applicable project policy, notice, and/or legal requirement).
- Transparency Principle: Projects involving PII will foster public trust by publishing PIAs and other public notices, except where the research is classified or Law Enforcement Sensitive (LES). PIAs will be conducted for classified or LES research projects, and when possible, a redacted version will be published.
- Redress Principle: The Privacy Office, in conjunction with S&T’s Privacy Officer, will develop and administer a redress program to handle inquiries and complaints regarding any S&T research projects involving PII.
- Training Principle: The Privacy office, in conjunction with S&T’s Privacy Officer, will provide privacy training for all project personnel regarding DHS privacy policy and any privacy protections specific to a particular project.